

# UNSUPERVISED UNDERSTANDING OF LOCATION AND ILLUMINATION CHANGES IN EGOCENTRIC VIDEOS

Alejandro Betancourt<sup>1,2</sup>  
a.betancourt@tue.nl

Natalia Díaz-Rodríguez<sup>3</sup>  
ndiaz@decsai.ugr.es

Emilia Barakova<sup>2</sup>  
e.i.barakova@tue.nl

Lucio Marcenaro<sup>1</sup>  
lucio.marcenaro@unige.it

Matthias Rauterberg<sup>2</sup>  
g.w.m.Rauterberg@tue.nl

Carlo Regazzoni<sup>1</sup>  
carlo@dibe.unige.it

<sup>1</sup> Department of Engineering (DITEN).  
University of Genova  
Genova, Italy

<sup>2</sup> Department of Industrial Design.  
Eindhoven University of Technology.  
Eindhoven, Netherlands.

<sup>3</sup> Computer Science Department.  
University of California Santa Cruz  
California, USA.

## ABSTRACT

Wearable cameras stand out as one of the most promising devices for the coming years, and as a consequence, the demand of computer algorithms to automatically understand these videos has been increasing quickly. An automatic understanding of these videos is not an easy task, and its mobile nature implies important challenges to be faced, such as the changing light conditions and the unrestricted locations recorded. This paper proposes an unsupervised strategy based on global features and manifold learning to endow wearable cameras with contextual information regarding the light conditions and the location recorded. Results show that non-linear manifold methods can capture contextual patterns from global features without compromising large computational resources. As an application case, the proposed unsupervised strategy is used as a switching mechanism to improve the hand-detection problem in egocentric videos under a multi-model approach.

*Index Terms*— Machine Learning, Unsupervised Learning, Egocentric Videos, First Person Vision, Wearable Camera

## 1. INTRODUCTION

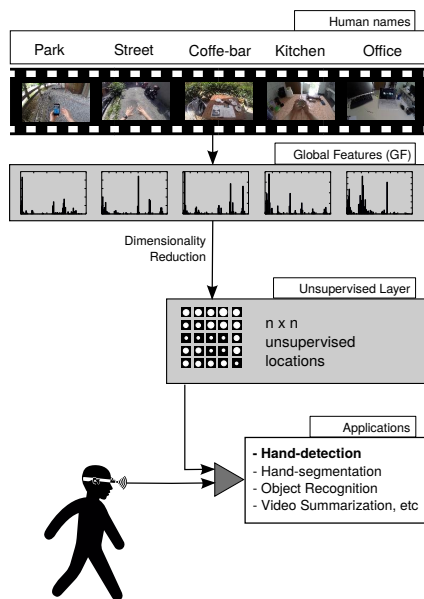
The emergence of wearable cameras such as action cameras, smart glasses and low-temporal life-logging cameras has detonated a recent trend in computer science know as First Person Vision (FPV) or Egovision. It is the first time that a camera is allowed to move with us in our daily activities and record what we see. The 90's idea of a wearable device with autonomous processing capabilities is nowadays possible and is considered one of the most relevant technological trends of the recent years [1]. The ubiquitous and personal nature of these devices opens the door to critical applications such as Activity Recognition[2], User-Machine Interaction[3], Ambient Assisting Living [4, 5, 6] Augmented Memory[7] and Blind Navigation [8], among others.

One of the key features of wearable cameras is their capability to move across different locations and record exactly what the user is looking at. Clearly an unrestricted video perspective that complicates the use of current FPV methods on real devices due to the unknown number of locations recorded from this video perspective. A common way to deal with this problem is to predefine a particular application or location and bound the algorithms based on this. It is the case of the gesture recognition for virtual museums proposed in [3] or the activity recognition methods based on the kitchen dataset [9, 10]. Another way to alleviate the large number of recorded locations is by using exhaustive video labeling of the recorded locations and objects as is done in [5] to detect daily activities. The authors in [11] use global histograms of color are used to reduce the effect of light changes in a color-based hand-segmenter.

The approach of [11] shows that contextual variables, such as light conditions, are valuable sources of information that can be used to improve the performance and applicability of current FPV methods. This idea is also applicable to other FPV applications such as activity recognition, on which a device that can understand the location of the user can easily reduce the possible activities and take more accurate decisions. Pervasive computing refers to the devices that can modify their behavior based on contextual variables as context-aware devices [12], and its benefits are widely explored for example in assisted living [13] and anomaly detection [14].

This paper is motivated by the potential impact of contextual information, such as light conditions and location, on different FPV methods. The strategy presented, is a first step towards our envision of a device that can understand the

environment of the user and modify its behavior accordingly. This paper understands the environment of the user as a set of different characteristics that can point to previously recorded conditions, and not as a scene classification problem based on manual labels assigned to particular locations (e.g., kitchen, office, street). In this way, The purpose of this study is to devise an unsupervised procedure for wearable cameras that can be used to switch between different models or search spaces according to the light conditions or location on which the user is involved. Figure 1 summarizes our approach.



**Fig. 1:** Unsupervised strategy to extract contextual information about light and location using global features.

As evident from Figure 1 the transition from the global features to the unsupervised layer can be understood as a dimensional reduction from the global feature space (high dimensional space) to a simplified low dimensional space (intrinsic dimension). The latter provides an unsupervised location map that can be used later to switch between different behaviours at different hierarchical levels. These dimensional reductions are known as manifold methods, and their capabilities to capture complex patterns are defined by their algorithmic and/or theoretic formulation [15].

Regarding the global features to be used, relevant information can be obtained from recent advances in FPV [1] and scene recognition [16, 17]. Given the restricted computational resources of wearable devices, we use computationally efficient features such as color histograms and GIST. However,

the proposed approach can be extended with more complex features such as deep features [14]. In that case three important issues must be considered: i) The computational cost will restrict the applicability in wearable devices. ii) It will require large amounts of training videos and manual labels. iii) The use of existent “pre-trained” neuronal architectures compromise the unsupervised nature of our approach.

The novelties of this paper are three folded: i) It evaluates the capability of different linear and non-linear manifold methods, namely Principal Component Analysis (PCA), Isometric Mapping (Isomaps) and Self Organizing Maps (SOM), to capture light/location patterns from different global features without using manual labels. ii) It analyzes, following a feature selection procedure, the most discriminative components of the selected global features, iii) As an application case, the proposed unsupervised strategy is used to improve the *hand-detection* problem in FPV. The hand-detection problem is used as an example, because of its impact on context-aware devices in hand-based methods, and because it allows us to illustrate the role of the unsupervised layer and its contribution to the final hand-detection performance. The use of the same strategy at higher inference levels such as hand-segmentation or hand-tracking is left as future research.

The remainder of this paper is organized as follows: Section 2 summarizes some recent strategies to understand automatically contextual information. Later, Section 3 introduces our methodological approach, summing up the selected features, different manifold methods and some common unsupervised evaluation procedures. In Section 4 the manifold methods are trained, and their capability to capture light/location patterns is evaluated in a post-learning strategy using the manual labels of two public FPV datasets. Section 5 illustrates the use of the best performing manifold method to improve the hand-detection rate in FPV. Finally, Section 6 concludes and provides some future research lines.

## 2. STATE OF THE ART

During the recent years, FPV video analysis is attracting the interest of the research community. The availability of wearable devices that can record what the user is looking at is ineluctable, and promising applications are emerging. Existing literature and commercial approaches highlights a broad range of possibilities, but also points to several challenges to be faced such as uncontrolled locations, illumination changes, camera motion, object occlusions, processing capabilities, among others [1]. This paper targets the issue of illumination changes as well as unrestricted locations recorded by the camera. The general idea is to develop an unsupervised layer that, based on global features and using low computational

resources, understands contextual information regarding the light conditions and the locations recorded by the camera.

The advantages of a device that can understand the environment are evident [18]. Recent advances in pervasive computing and wearable devices frequently point at the location of the user as a valuable information source to design context-aware systems [13, 12]. An intuitive way to find the location is to use Global Positioning Systems (GPS). However, this approach is commonly restricted by the battery life as well as by poor indoor signal [19].

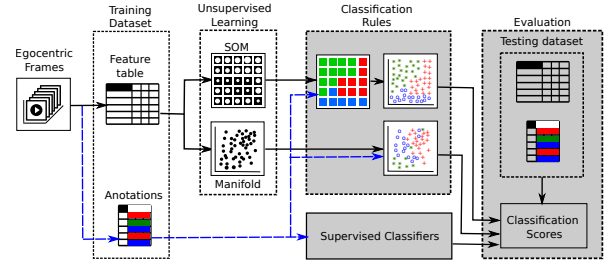
To alleviate these restrictions, wearable cameras, emerge as a possible solution: infer the context using the recorded frames. As an example, in [20] local and global features are combined to identify private locations and avoid recording them. In fact, the idea pursued by the authors is in line with the seminal works on scene recognition proposed by Oliva and Torralba, on which scenes captured by static cameras are represented as low dimensional vectors known as GIST [21, 17] and classified in a supervised way. Recent advances in scene recognition made by the same authors using deep features are promising [16]. However, their applicability on wearable devices is still restricted by the required computational resources and by the not-availability of large datasets recorded with wearable cameras.

Similar applications but following an unsupervised strategy are common in robotics, on which manifold algorithms like SOM or Neural Gas, are frequently used in autonomous navigation systems [22, 23, 24]. Regarding FPV, the authors in [11] propose multi-model recommendation system for hand-segmentation in egocentric videos that modify its internal behaviour based on the light conditions recorded. In their paper, the authors design a performance matrix containing one row per training frame and one column per model. The matrix values are the segmentation scores and are used to decide the model is the more suitable for each frame in the testing dataset.

Our proposed method is motivated by the switching mechanisms developed by [11]; However, it is independent on the segmentation dataset and can extract information about the light conditions as well as the recorded location. Regarding the scene-recognition literature, our approach is fully unsupervised and is based on computationally global features which make feasible to use it on wearable cameras.

### 3. OUR APPROACH

As explained in previous sections one of our goals is to quantify the capability of different unsupervised manifold methods to capture the illumination and location changes in egocentric videos. Our approach follows the experimental



**Fig. 2:** General workflow of our unsupervised evaluation. White blocks correspond to the unsupervised-learning. The manually labeled data is used only on shaded blocks, which correspond to the post-learning evaluation.

findings of previous works, on which global features such as Color-Histograms and GIST are used to describe the general characteristics of the scene [11, 17]. Figure 2 summarizes our approach. In the left part are the feature extraction and the unsupervised training. The right part shows the post-learning evaluation. The manual labels are used only in the shaded blocks of the diagram. The rest of this section introduces the datasets, motivates the global features and manifold methods, and concludes explaining the hyperparameter selection and the post-learning analysis.

#### 3.1. Datasets

Our comparison of the manifold methods is based on two popular FPV datasets, namely EDSH and UNIGE-HANDS. The main criteria for the dataset selection are the number of locations, the manual labels, and the availability of illumination changes in the videos.

**EDSH:** Dataset proposed by [11] to train a pixel-by-pixel Hand-Segmenter in FPV. The dataset contains 8 different locations with changing light conditions recorded from a head-mounted camera with a resolution of 720p at a speed of 30 *fps*. The labels about location and light conditions are manually created. For the experimental results, EDSH1 is used for training and EDSH2 for testing.

**UNIGE-HANDS:** Dataset proposed by [25] as baseline for the hand-detection problem in FPV. The dataset is recorded in 5 different locations (1. Office, 2. Coffee Bar, 3. Kitchen, 4. Bench, 5. Street), and is recorded with a resolution of  $1280 \times 720$  *pixels* and 50 *fps*. The dataset provides the locations of the videos. Labels about indoor/outdoor information were manually created. In Section 4 the original training/testing split is used.

### 3.2. Feature selection

To represent the scene context we use color histograms and GIST. These features are widely accepted and used in the FPV literature, and its computational cost makes them suitable for wearable devices with highly restricted processing capabilities and battery life. As explained before, more complex features such as deep features can be used under the same framework, but different issues must be faced to reach a real applicability. We point the use of deep features a promising future work.

Due to its straightforward computation and intuitive interpretation, color histograms are probably the most used features in image classification [26]. The variety of color spaces such as RGB, HSV, YCbCr or LAB makes possible to exploit color patterns while alleviating potential illumination issues. In particular, HSV is based on the way humans perceive colors while LAB and YCbCr use one of the components for lightness and the remaining ones for the color intensity. In egocentric vision, [27] uses a mixture of color histograms and visual flow for *hand-segmentation*, while [3] combined HSV features, a Random Forest classifier and super-pixels for gesture recognition. Recently, Li and Kitani [11] analyzed the discriminative power of different color histograms with a Random Forest regressor. Existent FPV literature commonly points to HSV as the best color space to face the changing light conditions in egocentric videos [27, 11]. For the experimental results, we use color histograms of RGB, HSV, YCbCr and LAB.

Additionally, we use GIST [28] as a global scale descriptor. It captures texture information, orientation and the coarse spatial layout of the image. GIST can be combined with other local descriptors to detect accurately objects in the scene, and was initially combined with a simple one-level classification tree, as well as with the naïve Bayesian classifier. GIST descriptor has been successfully applied on large scale image retrieval and object recognition [17].

The final part of the experimental results, analyzes the discriminative power, regarding light and location, of the proposed global features under a feature selection procedure. The idea behind this experiment is to fuse the more discriminative components of each global feature to increase the contextual information available in the high-dimensional space, and as consequence improve the patterns captured by the manifold method. Our hypothesis is that a combination of color (histograms) and global characteristics such as texture and orientations (GIST) could work better if fused properly. For this purpose, all the proposed global features are merged and used with a Random Forest to solve the classification problems explained in Section 4. The feature importance of the Random Forest is used to build a hand-crafted feature with

the most discriminative dimensions. The information available in the hand-crafted feature is compared against the original ones.

### 3.3. Manifold learning

Manifold methods are mathematical or algorithmic procedures designed to move from a high dimensional space to a low dimensional one while preserving the most valuable information [15]. Manifold methods are widely used and its applicability is fully validated in several field such as robotics [22, 23, 24], crowd analysis [29] and Speech recognition [30], among others.

In general, the capability of manifold methods to deal with complex data is defined by their mathematical formulations and assumptions. Manifold methods are usually grouped according to two factors: i) If the dimensional mapping uses manual labels, then the method is supervised; otherwise, it is unsupervised. As an example, Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are supervised and unsupervised, respectively. ii) If the intrinsic dimensions are linear combinations of original space then it is linear; otherwise, it is non-linear. As an example, PCA is linear, and SOM is non-linear. Due to the final objective of this paper, the remaining part does not consider the supervised approaches such as LDA.

To find a well performed dimensional mapping, we use as baseline the Principal Component Analysis (PCA) algorithm, which is the most common linear manifold algorithm but usually fails to capture patterns in complex datasets. To capture complex patterns we use two non-linear manifold methods, namely Isomaps and SOM. These non-linear algorithms were chosen based on the advantages reported in previous studies [31, 32], and their capability to be applied to new observations not included in the training data. In our exploratory analysis t-SNE was also used; however, its original formulation can not be applied to data outside of the training dataset. Regarding SOM, our study is based on the original formulation to keep simple the interpretation and analysis of the results; however, the same approach can be extended using recent versions of SOM such as the Hybrid-SOM [33] or the Growing-SOM [34]. The latter is particularly interesting to add some adaptive capabilities to the unsupervised layer, allowing it to increase the number of neurons while the user visits new locations.

**Principal Components Analysis:** It is a linear technique to reduce data dimensionality by transforming the original data into a new set of variables that summarize the original data [31]. The new variables are the principal components, and are uncorrelated and ordered such that the  $k$ -th PC has the  $k$ -th largest variance among all PCs, and the  $k$ -th PC is

orthogonal to the first  $k-1$  PCs. The first few PCs capture the main variations in the dataset, while the last PCs capture the residual "noise" in data.

**Isomaps:** A non-linear dimensionality reduction algorithm proposed in [31] that learns the underlying global geometry of a dataset using local distances between the observations. In comparison with classical linear techniques, Isomaps can handle complex non-linear patterns such as those in human handwriting or face recognition in images. Isomaps combine the major algorithmic features of PCA and the Multidimensional-scaling computational efficiency, global optimality, and asymptotic convergence, which makes feasible its use in wearable cameras. The hyperparameter of Isomaps is the number of neighbors [35].

**Self Organizing Maps (SOM):** It is one of the most popular unsupervised neural networks. It was originally proposed to visualize large dimensional datasets [36] and easily find relevant information [37] on them. In summary, the SOM is a two layer neural network that learns a non-linear projection of a high dimensional space (input layer) to a regular discrete low-dimensional grid of neural units (output layer). The discrete nature of the output layer facilitates the visualization of the learned patterns and makes easy to find topological relations in the data.

The training phase of SOM relies on a competitive recursive process with a neighborhood function that acts as a smoothing kernel over the output layer [36]. Typically, for each training sample, the best matching unit (BMU) is selected by using the Euclidean distance and then its local neighborhood is updated to make it slightly similar to the training sample. The neighborhood definition depends on the output layer. In our case, we use a regular quadrangular grid, but future improvements can be achieved by using more complex topologies such as toroidal or spherical grids [38]. The hyperparameter of SOM is the number of output neurons.

### 3.4. Hyperparameters, classification rules, and post-learning evaluation

When evaluating manifold-methods the most challenging part is to quantify if the patterns learned are ruled by the phenomena under study. Previous studies usually follow two different strategies: The first one quantifies the information lost when moving the training dataset from the original space to the intrinsic dimension [39]. The second strategy uses the manual labels or human knowledge to analyze the intrinsic dimension (output space) in a post-learning analysis [35].

In our case, the information strategy is used to define the hyperparameters of the Isomap and the SOM. In particular, we use the reconstruction error to select the number of neighbors

of the Isomaps as proposed in [39], and the Topological Conservation Quality (TCQ) to define the number of output neurons of SOM [30]. In general, the TCQ measures the number of times that the SOM transformation breaks a contiguity in the input data. In the input space, we define as contiguous two consecutive frames. In the output space two neurons are contiguous if they share one border. Formally the TCQ is defined as (1), where  $Q$  is the number of training samples and  $u(x_q) = 1$  if the two closest neurons of an input vector  $x_q$  are contiguous in the output space, and  $u(x_q) = 0$  otherwise.

$$TCQ = \frac{\sum_{q=1}^Q u(x_q)}{Q} \quad (1)$$

Once defined the hyperparameters, we follow the post-learning analysis using the manual labels to quantify the performance of the proposed manifold methods. For this purpose, each manifold method is trained on each global feature and dataset. Then a classification analysis is performed using the manual labels and defining as reference scores two popular supervised classifiers, namely Support Vector Machine (SVM) and Random Forest (RF). It is noteworthy that the supervised classifiers are in an advantageous position because they are theoretically developed to exploit the differences among manual labels; however, the closer the score of the manifold methods to the classifiers score, the more related the patterns learned are with the phenomena measured by the manual labels.

To use the manifold methods as classifiers, we use a majority voting rule in the output space (intrinsic dimension) using the training samples and their manual labels. For Isomaps and PCA, the majority voting rule is evaluated using the 10 closest training frames in the output space. For SOM, the majority voting rule is evaluated on the training frames that activated the same output neuron of each testing sample.

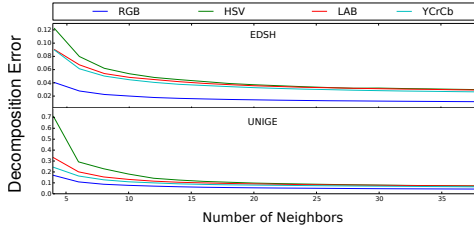
## 4. EXPERIMENTAL RESULTS

This section evaluates the capabilities of the proposed manifold methods to capture light changes and separate different locations using global features. In the first part of this section, we calibrate the hyperparameters of the Isomap and SOM while preserving the unsupervised nature of the training phase. Later, we use the manual labels to analyze the patterns learned under a classification approach [35]. Finally, the discriminative ranking learned by a Random Forest is used to analyze the most relevant dimensions of the proposed global features.

### 4.1. Defining the hyperparameters

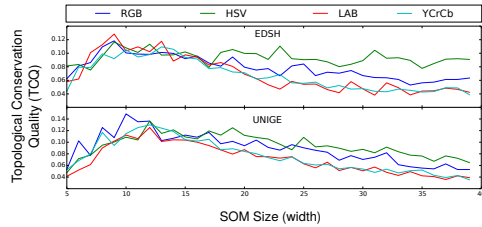
To define the number of neighbors considered in Isomaps we use the reconstruction error, which is the amount of

information lost when transforming a point from the original space (global feature) to the intrinsic dimension. Figure 3 shows the reconstruction error of the Isomap when the number of closest neighbors increases. Note that, for all the features; the reconstruction error starts stabilizing when the 12 closest neighbors are used. Therefore, we use 12 as the parameter in the remaining part of the paper.



**Fig. 3:** Isomap reconstruction error in function of the number of neighbors

Regarding the number of output neurons of the SOM we use the TCQ, as defined in Section 3.4. Figure 4 shows the TCQ for different SOM sizes. Two findings from the figure: i) A small number of neurons offers a topological advantage in the TCQ, because the fewer the output neurons to activate, the easier to preserve contiguities in the output space. ii) The TCQ starts stabilizing for large SOMs, around  $20 \times 20$  for EDSH and  $30 \times 30$  for UNIGE dataset. In the experimental results we use three SOM sizes:  $5 \times 5$ ,  $20 \times 20$  and  $30 \times 30$ , denoted as  $SOM_5$ ,  $SOM_{20}$ ,  $SOM_{30}$  respectively.



**Fig. 4:** TCQ error in function of the number of neighbors

#### 4.2. Post-learning analysis

To evaluate the patterns found by the manifold methods we perform an exhaustive post-learning analysis under a classification framework using the manual labels and defining as reference scores the performance of SVM and RF. For this purpose we define two different classification problems: i) Discriminate among indoors and outdoors frames ii) Classify the labeled locations given by the datasets (e.g. Kitchen, Office, Street, etc).

**Table 1:** Supervised evaluation of different methods (columns) when used on top of different features (rows). Performance values are presented in two horizontal groups, one per classification problem (indoor/outdoor and location). The performance of both datasets (EDSH and UNIGE-Hands) are presented. The values reported are accuracy (properly classified frames over the total testing frames) for each feature/method combination.

		Unsupervised					Supervised		
	dataset	feature	$SOM_5$	$SOM_{20}$	$SOM_{30}$	PCA	Isomap	SVM	RF
Indoor and Outdoor	EDSH	RGB	0.679	0.796	0.767	0.743	0.743	0.790	0.847
		HSV	0.775	0.774	<b>0.847</b>	0.731	0.752	<b>0.891</b>	0.859
		LAB	0.694	0.747	0.711	0.657	0.781	0.843	0.839
		YCrCb	0.616	0.618	0.627	0.627	0.615	0.763	0.791
		GIST	0.650	0.820	0.757	0.634	0.655	0.749	0.787
	UNIGE	RGB	0.897	0.925	0.911	0.605	0.665	0.917	0.972
		HSV	0.979	<b>0.988</b>	<b>0.986</b>	0.878	0.932	0.972	<b>0.987</b>
		LAB	0.972	0.967	0.959	0.770	0.964	0.981	0.989
		YCrCb	0.805	0.900	0.878	0.770	0.898	0.972	0.984
		GIST	0.753	0.857	0.834	0.672	0.735	0.962	0.896
Location	EDSH	RGB	0.340	0.488	0.474	0.440	0.474	0.503	0.630
		HSV	0.418	0.519	<b>0.621</b>	0.327	0.343	0.551	<b>0.671</b>
		LAB	0.352	0.371	0.366	0.284	0.389	0.452	0.570
		YCrCb	0.325	0.255	0.267	0.195	0.229	0.330	0.521
		GIST	0.462	0.526	0.519	0.285	0.367	0.554	0.518
	UNIGE	RGB	0.730	0.858	0.824	0.405	0.467	0.842	0.933
		HSV	0.826	0.957	<b>0.960</b>	0.671	0.826	0.953	<b>0.957</b>
		LAB	0.790	0.831	0.787	0.540	0.815	0.919	0.936
		YCrCb	0.669	0.797	0.752	0.622	0.792	0.919	0.922
		GIST	0.436	0.656	0.665	0.353	0.400	0.881	0.740

Table 1 shows the percentage of testing data successfully classified by each method (columns) when using different features (rows). The table contains two horizontal groups, one for each classification problem. The first group shows the performance for the binary problem (indoor/outdoor), and the second group shows the strict multiclass match for the detailed locations. The first group of columns shows the unsupervised methods while the second group shows the supervised classifiers results. Note that, despite not using manual labels in the training phase, the performance of the unsupervised methods are close to their supervised counterparts, which validates the patterns learned, and confirms the relationship between the proposed global features with the light/location conditions.

In particular, Table 1 shows that, within the unsupervised techniques, SOMs perform the best. Regarding the SOM size,  $SOM_{30}$  is the best choice for both datasets and classification problems. The table also shows valuable insights about the most discriminative features. It is noteworthy the performance of the methods when HSV is used, particularly in the unsupervised approach. This fact confirms the intuition of previous works on which the use of HSV leads to algorithmic improvements when used as a proxy for the light conditions. About the datasets, it is possible to conclude that the EDSH dataset is the most challenging, especially for the location classification problem.

Another intuitive way to analyze the results is by visualizing the patterns learned. In summary, a well performed dimen-

sional mapping must locate frames close to each other, in the output space, if they are under similar light conditions and scene configuration. In other words, if the proposed features are related to the light/location conditions, the unsupervised method will try to separate them in the output space. The quality of that separation is ruled by the complexity of the data and the manifold method used.

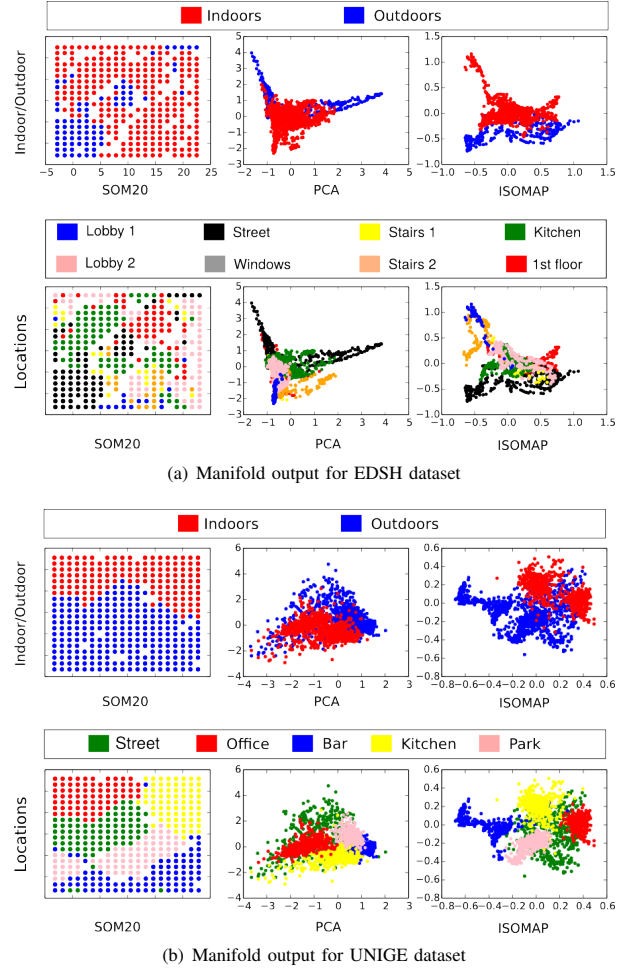
Figure 5 shows the 2D output for the  $SOM_{20}$ ,  $Isomap$ , and  $PCA$ , for both datasets using HSV. Different colors represent the manual labels. In the case of SOM, each neuron is labeled with the majority voting of the neuronal activations hits. The figure clearly shows that SOM successfully group similar inputs in the same regions of the output layer. In the case of PCA and Isomaps, the patterns in the output space are not so evident, but definitely, the non-linearity of Isomaps allows them to capture more information than PCA, which is clearly affected by the orthogonality of the intrinsic dimensions.

It is remarkable the output space of the  $SOM_{30}$  in the UNIGE dataset, on which both classification problems are located in different parts of the output layer. For the EDSH dataset, it is also possible to delineate some clusters, such as the kitchen (green) the street (black), the 1st floor (red) and the stairs (yellow and orange). However, the remaining locations are not easily visible, e.g., both lobbies (in blue and pink). This is explained by the small number of frames available for these locations in the dataset. This problem can be faced by using dynamic versions of SOM like Growing-SOM [35].

Finally, Figure 6 shows the  $SOM_{30}$  signature when transforming a uniform sampling of 40 seconds from the street video of the UNIGE dataset using HSV. In the first row are the activated neurons (unsupervised locations) ordered by time from left to right. In the second row are the compressed snapshots for the input frames. As can be seen from the first row, the  $SOM_{30}$  activations start on the left side and moves to the middle of the grid while the user walks in the street through different light conditions. The point color represents the temporal dimension, being yellow the first frame and red the last one.

#### 4.3. Hand-crafted feature

This subsection exhaustively analyzes the discriminative capabilities of the proposed global features and combine the most relevant dimensions to improve the dimensional mapping. For this purpose we follow two steps: i) The global features (RGB, HSV, LAB, YCrCb, GIST) are combined and used to train a RF on each dataset and classification problem described in section 4. ii) The discriminative importance learned by the RF is exploited by adding, in order of importance, each of the original dimensions while evaluating the performance of RF and  $SOM_{30}$ .



**Fig. 5:** 2D representation of the datasets using SOM, PCA, and Isomaps, for the EDSH (a) and UNIGE (b) datasets.

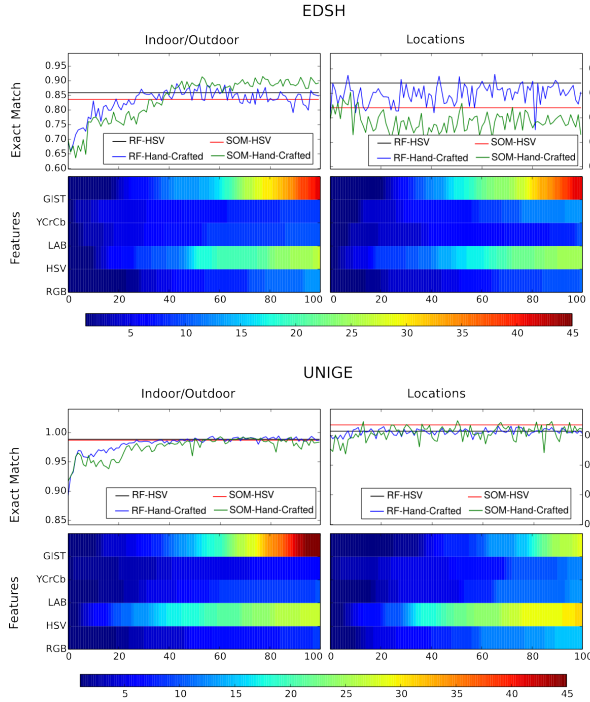
Figure 7 summarizes the changes in performance (line plot) and the number of components (heat-map) belonging to each global feature on each step (x-axis). The upper and lower parts of the figure show the results for the EDSH and the UNIGE dataset, respectively. The first column corresponds to the indoor/outdoor problem and the second column to the location problem. The constant values in the line plots are the performance of  $SOM_{30} - HSV$  and  $RF - HSV$  reported in Table 1.

From Figure 7 it is possible to conclude that hand-crafted features could improve the performance in the proposed classification problems. For instance, for the EDSH dataset,





**Fig. 6:**  $SOM_{30}$  signature for 40 seconds from the street video in the UNIGE dataset. The first row shows the activated neurons in the SOM output layer by the frame presented in the lower row.



**Fig. 7:** Performance of hand-crafted features by adding on component by step: **top:** EDSH dataset **bottom:** UNIGE dataset. First column shows the indoor/outdoor problem and second column visualizes the location problem (Section 4). The lines plot represents performance and the heat maps represent the number of components selected in each step from each original feature. The color bars below the heat maps show the legend relating a color with a particular number.

the hand-crafted feature improves the SOM accuracy from 84.7% to 91.4% and 62.1% to 65.2% in the indoor/outdoor and location problem, respectively. For the UNIGE dataset, due to the original performance, the improvement is not as significant. However, for some steps in the location problem, the hand-crafted feature reaches an accuracy of 99.2%, which is slightly better than the 98.7% of the HSV version. It is also noteworthy the result on the location problem for the EDSH dataset, on which the hand-crafted feature is close to the SOM-

HSV combination, but is not able to improve the performance considerably. The latter fact confirms that the location problem in the EDSH dataset is one the most challenging, not only for the manifold methods but also for the supervised classifiers.

Regarding the composition of the hand-crafted feature, it is notable that by using less than 40 components, it is possible to achieve similar performance to the SOM-HSV, which originally uses 94 components. Additionally, for all cases, the method starts using HSV components as the most discriminative, but around the 30 to the 40 step, it aggressively uses GIST components to disambiguate the most difficult cases. A quick analysis of the GIST components suggests that the RF searches for orientations and scale in the scene. Finally, the other color-spaces are barely used. Table 2 shows the confusion matrix of SOM and RF using the global HSV and the hand-crafted feature for the location problem of the UNIGE dataset. Confusion scores indicate that the office and park locations are the most difficult to discriminate.

**Table 2:** Confusion matrix of SOM and RF using the global HSV and the hand-crafted feature for the location problem of UNIGE dataset. Values represent percentages.

		SOM					RF				
		Kitchen	Office	Bar	Street	Park	Kitchen	Office	Bar	Street	Park
HSV	Kitchen	<b>98.4</b>	0.3	0.3	0.0	0.9	<b>97.8</b>	0.0	0.3	0.0	1.9
	Office	1.8	<b>92.4</b>	2.5	0.0	3.3	0.4	<b>86.6</b>	1.1	0.0	12.0
	Bar	0.0	0.0	<b>99.7</b>	0.0	0.3	0.0	0.0	<b>99.0</b>	0.7	0.3
	Street	0.4	0.0	0.4	<b>98.5</b>	0.7	0.0	0.0	0.7	<b>98.9</b>	0.4
	Park	0.8	2.9	1.3	0.8	<b>94.2</b>	1.3	2.5	0.0	0.4	<b>95.8</b>
Hand Crafted	Kitchen	<b>98.4</b>	0.0	0.3	0.0	1.2	<b>97.8</b>	0.3	0.0	0.0	1.9
	Office	0.0	<b>94.6</b>	0.4	0.0	5.1	0.4	<b>86.6</b>	1.1	0.0	12.0
	Bar	0.0	0.0	<b>99.7</b>	0.0	0.3	0.0	0.0	<b>99.7</b>	0.3	0.0
	Street	0.4	0.0	1.8	<b>97.8</b>	0.0	0.0	0.0	1.5	<b>96.7</b>	1.8
	Park	1.3	1.3	0.4	0.4	<b>96.7</b>	2.1	2.1	0.0	0.0	<b>95.8</b>

## 5. APPLICATION CASE: MULTI-MODEL HAND-DETECTION

Once confirmed the capabilities of SOM to capture light conditions and the global characteristics of the scene, its output can be used as a map of unsupervised locations to



build a multi-model approach to different problems such as object recognition, hand-detection, video-summarization, activity recognition, among others. This section illustrates the use of the unsupervised layer by using the hand-detection problem as defined in [40], on which a Support Vector Machine (SVM) is trained with Histogram of Oriented Gradients (HOG) to detect whether the hands are being recorded by the camera or not [40, 41].

The hand-detection problem is used as example due to two reasons: 1) It solves a simple question that allows us to illustrate the workflow and the role of the unsupervised layer in the reported improvements 2) The manual labeling is simple and easy to replicate. A similar approach can be extended to other hierarchical levels (e.g. hand-segmentation); However, it would require extra labeling to supply the number of neurons that grows quadratic.

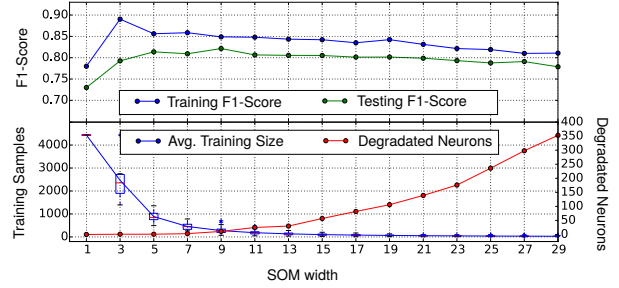
Our approach extends the method proposed in [40] by training one hand-detector for each unsupervised neuron of the HSV-SOM described in Section 3. For each neuron  $i \in SOM_N$  we denote its localized hand-detectors as  $hd_i^N$ , and the global hand-detector as  $hd^N$ . Given an arbitrary frame  $f$ , the local and global confidence about the hand presence is given by the SVM probabilistic notation as stated in equation (2) and (3), respectively. Here  $\Theta$  refers to the hyperplane learned by the HOG-SVM when trained on the whole training videos and  $\theta_i$  the hyperplane learned when using for training only on the activations of the neuron  $i$ . For practical purposes, we use each training frame to train its 5 best matching neurons.

$$hd_i^N(f) = SVM(HOG(f)|\theta_i) \quad (2)$$

$$hd^N(f) = SVM(HOG(f)|\Theta) \quad (3)$$

Regarding the training set for each neuron, some facts must be taken into account. Due to the finite number of training frames, some neurons does not have enough training activations. We denote these neurons as degraded. Additionally, the training activations of the local neighborhood of each neuron  $i$  can be used to construct a local testing set (LTS) for  $hd_i^N$ , and using it to estimate its local  $f1$ . We define the neurons with local  $f1$  lower than 0.75 as degraded. The confidence of the degraded neurons is set to 0 and the global hand-detector is used in case of activation.

Figure 8 summarizes the performance of the multi-model approach for different SOM sizes (x-axis). The upper half of the figure shows the training and testing  $F1$  scores. This figure shows a quick increase in the  $F1$  score which stabilize for SOMs with more than  $9^2$  neurons. The lower half of the figure shows the average number of training frames per neuron (blue)



**Fig. 8:** The upper part of the figure shows the training (blue) and testing (green)  $F1$  score. The lower part shows the average number of training frames (blue) used to train  $hd_i^N \in SOM_N$ , and the number of degraded neurons (red). The horizontal axis is the size of the SOM.

**Table 3:** True-Positives and True-Negatives comparison between a unique model approach as proposed in [40] (HOG-SVM) and a Multimodel approach using the  $SOM_9$  (Ours)

	True positive rate		True negatives rate		F1-score	
	HOG-SVM	Ours	HOG-SVM	Ours	HOG-SVM	Ours
Office	0.888	<b>0.914</b>	0.928	<b>0.937</b>	0.897	<b>0.917</b>
Street	0.767	<b>0.797</b>	0.871	<b>0.927</b>	0.814	<b>0.856</b>
Bench	0.743	<b>0.799</b>	0.964	<b>0.966</b>	0.832	<b>0.868</b>
Kitchen	0.618	<b>0.646</b>	0.773	<b>0.794</b>	0.691	<b>0.718</b>
Coffee bar	0.730	<b>0.805</b>	0.695	<b>0.767</b>	0.718	<b>0.790</b>
Total	0.739	<b>0.783</b>	0.846	<b>0.877</b>	0.780	<b>0.821</b>

and the number of degraded neurons (red). Two important conclusions from these figures: i) The multi-model approach overfit the training dataset on large SOMs ii) The number of degraded neurons increases quickly and as the consequence no extra benefit is obtained from larger SOMs. These findings reinforce the idea of using a dynamic self-organizing structure in future research.

Table 3 compares the performance of the HOG-SVM and the multi-model strategy on a  $SOM_9$ . The table shows the true-positive rate, true-negative rate and the  $F1$  score for each location in the dataset. In general, our approach considerably improves the performance for all locations, totalizing an improve of 4.1  $F1$  points in the whole dataset. The location with the larger improvement is the *Coffe-bar* with an increase of 7 points in the  $F1$ . This improvement is explained by an increase of 7.5 and 7.2 percentual units in the true-positive and true-negative rate respectively.

Finally, Figure 9 summarizes some neuronal characteristics of the  $SOM_9$ : Figure 9(a) shows the number of training frames used per neuron and the proportions between frames with (green) and without (red) hands. Some neurons have a slightly unbalanced training. This fact is included in the hand-detector training face by using these proportions as the weights

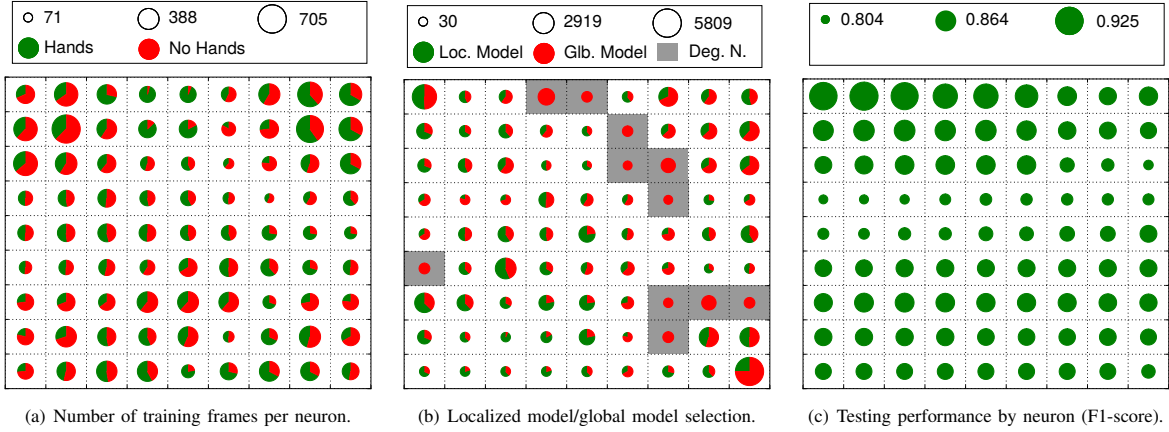


Fig. 9: Important facts about the multimodel approach when using  $SOM_9$  as model based.

of the class in the SVM. Figure 9(b) summarizes the use of  $hd_t^9$  and  $hd^9$ . The size of the circle represents the number of testing frames activating a particular neuron. In turn, each circle is proportionally divided in green and red according to the number of times that the local or global model is used, respectively. The gray cells are the degraded neurons on which only the global model is used. Finally, the Figure 9(c) shows the testing F1 score of each neuron. It is noteworthy that the smallest  $F1$  scores are located in a contiguous region of the  $SOM_9$ . This fact can be exploited by using a windowing to fuse the local models. In sake of an easy explanation of the application case this improvement is not included in the current implementation.

## 6. CONCLUSIONS AND FUTURE RESEARCH

This paper proposes an unsupervised strategy to endow wearable cameras with contextual information about the light conditions and location recorded by using global features. The main finding of our approach is that using SOM and HSV, it is possible to develop an unsupervised layer that understands the illumination and location characteristics on which the user is involved. Our experiments validate the intuitive finding of previous works on which HSV global histograms are used as a proxy for the light conditions recorded by a wearable camera. As an application case, the unsupervised layer is used to face the hand-detection problem under a multi-model approach. The experiments presented in the hand-detection application considerably outperform the method proposed in [40].

The experimental results analyze the capabilities of different unsupervised methods to capture light and location changes in egocentric videos. The experimental results show that

SOM can extract valuable contextual information about the illumination and location from egocentric videos without using manually labeled data.

Regarding the relationship between the global features and the recorded characteristics, our experiment points at HSV as the color space having the most discriminative power. Additionally, it is shown that by following a simple feature selection, it is possible to develop a hand-crafted feature, mainly formed by HSV and GIST, which makes easier for SOM to capture these patterns. Two issues about the hand-crafted feature to be accounted for: i) it is computationally expensive compared with using just HSV. ii) It indirectly introduces a dependence between the manual labels and the training phase.

Concerning future work, several challenges in the proposed method can be faced. One of the more promising is the use of deep features to extract more complex contextual patterns. Another interesting research line is the use of dynamic versions of SOM such as the Growing-SOM [34], to allow the unsupervised layer to constantly change its output space while the user visit new locations. In the application case, important improvements can be achieved if the proposed framework is applied to other hierarchical levels, for example, the unsupervised layer can be used to switch between different color spaces at a hand-segmentation level or used to select different dynamic models at a hand-tracking level.

Finally, an interesting application of the proposed approach can be found in video summarization, visualization and captioning. In this line, the output space can be used to find easily and retrieve video segments recorded on similar locations or light conditions.

## 7. ACKNOWLEDGMENT

This work was partially supported by the Erasmus Mundus joint Doctorate in Interactive and Cognitive Environments, which is funded by the EACEA, Agency of the European Commission under EMJD ICE. Likewise, we thank the AAPELE (Architectures, Algorithms and Platforms for Enhanced Living Environments) EU COST action IC1303 for the STSM Grant, the International Neuroinformatics Coordinating Facility (INCF) and the Finnish Foundation for Technology Promotion (TES).

## REFERENCES

- [1] A. Betancourt, P. Morerio, C. Regazzoni, and M. Rauterberg, "The Evolution of First Person Vision Methods: A Survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 744–760, 2015.
- [2] T.-H.-C. Nguyen, J.-C. Nebel, and F. Florez-Revuelta, "Recognition of Activities of Daily Living with Egocentric Vision: A Review," *Sensors*, vol. 16, no. 1, pp. 72, 2016.
- [3] L. Baraldi, F. Paci, G. Serra, L. Benini, and R. Cucchiara, "Gesture Recognition using Wearable Vision Sensors to Enhance Visitors' Museum Experiences," *IEEE Sensors Journal*, vol. 15, no. 5, pp. 1–1, 2015.
- [4] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *Proceedings of the IEEE International Conference on Computer Vision*, nov 2011, pp. 407–414, IEEE.
- [5] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, jun 2012, pp. 2847–2854, IEEE.
- [6] N. Díaz, M. Pegalajar, J. Lilius, and M. Delgado, "A survey on ontologies for human behavior recognition," *ACM Computing Surveys*, vol. 46, no. 4, pp. 1–32, 2014.
- [7] J. Farrington and V. Oni, "Visual Augmented Memory," in *International Symposium on wearable computers*, Atlanta GA, 2000, pp. 167–168.
- [8] G. Balakrishnan, G. Sainarayanan, R. Nagarajan, and S. Yaacob, "Wearable Real-Time Stereo Vision for the Visually Impaired," *Engineering Letters*, vol. 14, no. 2, pp. 6–14, 2007.
- [9] A. Fathi and J. M. Rehg, "Modeling actions through state changes," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, jun 2013, pp. 2579–2586, Ieee.
- [10] K. Matsuo, K. Yamada, S. Ueno, and S. Naito, "An Attention-Based Activity Recognition for Egocentric Video," in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, jun 2014, pp. 565–570, Ieee.
- [11] C. Li and K. M. Kitani, "Model Recommendation with Virtual Probes for Egocentric Hand Detection," in *2013 IEEE International Conference on Computer Vision*, Sydney, 2013, pp. 2624–2631, IEEE Computer Society.
- [12] O. D. Lara and M. A. Labrador, "A Survey on Human Activity Recognition using Wearable Sensors," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.
- [13] D. Riboni and C. Bettini, "COSAR: hybrid reasoning for context-aware activity recognition," *Personal and Ubiquitous Computing*, vol. 15, no. 3, pp. 271–289, 2011.
- [14] Y. Zhu, N. M. Nayak, and a. K. Roy-Chowdhury, "Context-Aware Activity Recognition and Anomaly Detection in Video," *Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 91–101, 2013.
- [15] J. H. Friedman, "On bias, variance, 0/1-loss, and the curse-of-dimensionality," *Data mining and knowledge discovery*, vol. 77, no. 1, pp. 55–77, 1997.
- [16] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning Deep Features for Scene Recognition using Places Database," *Advances in Neural Information Processing Systems* 27, pp. 487–495, 2014.
- [17] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.
- [18] T. Starner, B. Schiele, and A. Pentland, "Visual contextual awareness in wearable computing," in *Digest of Papers Second International Symposium on Wearable Computers Cat No98EX215*, 1998, pp. 50–57, IEEE Computer Society.
- [19] T. Hori and K. Aizawa, "Context-based video retrieval system for the life-log applications," in *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval - MIR '03*, New York, New York, USA, 2003, p. 31, ACM Press.
- [20] R. Templeman, M. Korayem, D. Crandall, and K. Apu, "PlaceAvider: Steering first-person cameras away from sensitive spaces," in *Network and Distributed System Security Symposium*, 2014, number February, pp. 23–26.
- [21] A. Oliva, "Gist of the scene," in *Neurobiology of Attention*, pp. 251–256, Elsevier Inc., 2005.
- [22] P. Baldassarri and P. Puliti, "Self-organizing maps versus growing neural gas in a robotic application," in *Artificial Neural Nets Problem Solving Methods*, pp. 201–208, 2003.
- [23] E. Barakova and T. Lourens, "Event Based Self-Supervised Temporal Integration for Multimodal Sensor Data," *Journal of Integrative Neuroscience*, vol. 04, no. 02, pp. 265–282, 2005.
- [24] E. Barakova and T. Lourens, "Efficient episode encoding for spatial navigation," *International Journal of Systems Science*, vol. 36, no. October 2014, pp. 887–895, 2005.
- [25] A. Betancourt, P. Morerio, E. Barakova, L. Marcenaro, M. Rauterberg, and C. Regazzoni, "A Dynamic Approach and a New Dataset for Hand-Detection in First Person Vision," in *International Conference on Computer Analysis of Images and Patterns*, Malta, 2015.
- [26] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," *Pattern Recognition*, vol. 40, no. 3, pp. 1106–1122, mar 2007.
- [27] P. Morerio, L. Marcenaro, and C. Regazzoni, "Hand Detection in First Person Vision," in *Isip40.It*, Istanbul, 2013, University of Genoa, pp. 0–6.
- [28] K. Murphy, A. Torralba, D. Eaton, and W. Freeman, "Object detection and localization using local and global features," *Toward Category-Level Object Recognition - Lecture Notes in Computer Science*, vol. 4170, pp. 382–400, 2006.
- [29] S. Chiappino, P. Morerio, L. Marcenaro, and C. S. Regazzoni, "Bio-inspired relevant interaction modelling in cognitive crowd management," *Journal of Ambient Intelligence and Humanized Computing*, vol. Feb, no. 1, pp. 1–22, feb 2014.
- [30] N. Arous and N. Ellouze, "On the Search of Organization Measures for a Kohonen Map Case Study: Speech Signal Recognition," *International Journal of Digital Content Technology and its Applications*, vol. 4, no. 3, pp. 75–84, 2010.
- [31] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
- [32] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.
- [33] M. O. Afolabi and O. Olude, "Predicting stock prices using a hybrid Kohonen Self Organizing Map (SOM)," in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2007, pp. 1–8.
- [34] G. Zhu and X. Zhu, "The growing self-organizing map for clustering algorithms in programming codes," in *International Conference on Artificial Intelligence and Computational Intelligence, AICI 2010*, 2010, vol. 3, pp. 178–182.
- [35] L. Jing and C. Shao, "Selection of the suitable parameter value for ISOMAP," *Journal of Software*, vol. 6, no. 6, pp. 1034–1041, 2011.
- [36] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, Sep 1998.
- [37] T. Kohonen, "Self-organizing maps," *Springer Series in Information Sciences. Berlin, Heidelberg*, vol. 30, no. 3rd edition, 2001.
- [38] N. J. Mount and D. Weaver, "Self-organizing maps and boundary effects: Quantifying the benefits of torus wrapping for mapping SOM trajectories," *Pattern Analysis and Applications*, vol. 14, no. 2, pp. 139–148, 2011.
- [39] A. Saxena and A. Gupta, "Non-linear dimensionality reduction by locally linear isomaps," *Neural Information Processing*, pp. 1038–1043, 2004.
- [40] A. Betancourt, P. Morerio, L. Marcenaro, M. Rauterberg, and C. Regaz-

zoni, "Filtering SVM frame-by-frame binary classification in a detection framework," in *International Conference on Image Processing*, Quebec, Canada, 2015, IEEE.

- [41] A. Betancourt, M. Lopez, C. Regazzoni, and M. Rauterberg, "A Sequential Classifier for Hand Detection in the Framework of Egocentric Vision," in *Conference on Computer Vision and Pattern Recognition*, Columbus, Ohio, jun 2014, vol. 1, pp. 600–605, IEEE.



**Alejandro Betancourt** Alejandro Betancourt is PhD candidate of the Interactive and Cognitive Environment program between the Università degli Studi di Genova and the Eindhoven University of Technology. Alejandro is a Mathematical Engineer and Master In Applied Mathematics from EAFIT University (Medellin, Colombia). Since 2011 Alejandro has been involved in research about Artificial Intelligence, Machine Learning and Cognitive Systems.



**Natalia Díaz-Rodríguez** Natalia Díaz-Rodríguez received her double co-supervised PhD on Computer Engineering from Åbo Akademi University (Finland) and University of Granada (Spain) in 2015 (cum laude) and is currently research fellow at University of California Santa Cruz, Computer Science Department, at the Statistical Relational Learning (LINQS) group. Her research interests range, in Artificial Intelligence, from machine and deep learning to robotics, semantics, wearables, data science and other scalable statistical learning methods for every-day problems.



**Emilia Barakova** Emilia I. Barakova received the Master's degree in electronics and automation from the Technical University of Sofia, Bulgaria, and the Ph.D. degree in mathematics and physics from Groningen University, The Netherlands, in 1999. She is currently with the Department of Industrial Design, Eindhoven University of Technology, The Netherlands. She has expertise in Artificial Intelligence, Robotics, and User-Centered Design and has been affiliated with Fraunhofer AiS, GMD-Japan and RIKEN Brain Science Institute, Japan.



**Lucio Marcenaro** Lucio Marcenaro received his PhD in Computer Science and Electronic Engineering from University of Genova in 2003. From 2003 to 2010 he was CEO and development manager at TechnoAware srl. From March 2011, he became Assistant Professor in Telecommunications at the University of Genova. He is the principal scientific and technical coordinator of the Ambient Awareness Lab (A2Lab), with TechnoAware srl. He authored over 30 technical papers related to signal and video processing for computer vision. His main current research interests are: video processing for event recognition, detection and localization of objects in complex scenes, distributed heterogeneous sensors ambient awareness systems, ambient intelligence and bio-inspired cognitive systems.



**Matthias Rauterberg** Matthias Rauterberg is professor at the department of Industrial Design and the head of the Designed Intelligence group at Eindhoven University of Technology (The Netherlands). Matthias received the B.S. in Psychology (1978) at the University of Marburg (Germany), the B.S. in Philosophy (1981) and Computer Science (1983), the M.S. in Psychology (1981, summa cum laude) and Computer Science (1986, summa cum laude) at the University of Hamburg (Germany), and the Ph.D. in Computer Science/Mathematics (1995, awarded) at the University of Zurich (Switzerland).



**Carlo Regazzoni** Carlo S. Regazzoni received the Laurea degree in Electronic Engineering and the Ph.D. in Telecommunications and Signal Processing from the University of Genoa (UniGE), in 1987 and 1992, respectively. Since 2005 Carlo is Full Professor of Telecommunications Systems. Dr. Regazzoni is involved in research on Signal and Video processing and Data Fusion in Cognitive Telecommunication Systems since 1988.